CLARIN
Common Language Resources and Technologies Infrastructure

Working rapporteur: Steven Krauwer

The CLARIN project is a large-scale pan-European collaborative effort to create, coordinate and make language resources and technology available and readily useable to scholars of all disciplines, in particular the humanities and social sciences.

It intends to rise to the challenge that language (in spoken, written, multimodal form, as carrier of information, object of study, or otherwise) presents in our age when the volume of texts (either as continuous discourse or, for example, descriptions of objects of cultural heritage) and (more recently) recorded spoken texts is enormous, and it is growing exponentially. The sheer size of this material makes the use of computer-aided methods indispensable for many scholars in the humanities and in neighbouring areas who are concerned with language material.

The CLARIN infrastructure is based on the firm belief that the days of pencil-and-paper research are numbered even in the Humanities. Computer aided language processing is already used by a wide variety of sub-disciplines in the humanities and social sciences, addressing one or more of the multiple roles language plays (i.e. carrier of cultural content and knowledge, instrument of communication, component of identity and object of study). There is a high degree of commonality in the methods and objectives of current practice, and it is also evident that to reach the higher levels of analysis of texts which non-linguist scholars are typically interested in, such as their semantic and pragmatic dimensions, requires an effort of a scale that no single scholar could, or indeed, should afford.

The cost of collecting, digitising and annotating large text or speech corpora, dictionaries or language descriptions is huge in terms of time and money, and the creation of tools to manipulate these language data is very demanding in terms of skills and expertise, especially if one wants to make them accessible to professionals who are not experts in linguistics or language technology.

Hence the benefits of computer enhanced language processing become available only when a critical mass of coordinated effort is invested in building an enabling infrastructure, which can then provide services in the form of provision of tools and resources as well as training and counselling across a wide span of domains. This is the mission of the CLARIN infrastructure initiative.

To realize the above objectives, CLARIN will create a comprehensive and free to use archive of language resources and technologies covering not only the languages of all member states, but also minority languages and language phenomena addressing the issue of migration. The infrastructure will be based on a number of resource, service and expertise centres and will commit itself to collaborating with education organizations and to providing training programs with the aim of enabling the widest possible range of users to exploit the benefits in their own field. Through the fact that the tools and resources will be interoperable across languages and domains will in itself contribute a great deal towards addressing the issue of preserving and supporting multilingual and multicultural European heritage. An operational open infrastructure of web services will introduce a new paradigm of distributed collaborative development. It will allow many contributors to add all kinds of new services based on existing ones thus ensuring reusability and allowing scaling up to suit individual needs.

The CLARIN community unites all the leading institutions in the Language Resources and Technologies field across Europe representing decades of experience in tools and resources development, standardisation and infrastructure initiatives. The existing networks (ELRA, ELSNET, TELRI etc.) are now set to join forces in this truly pan-European initiative. The governance and management plans for the CLARIN infrastructure are detailed in a separate document.

Regarding the potential users of the CLARIN infrastructure, we should emphasize that the overall objective of the infrastructure is to bring computational analysis of texts and semantic annotation within the means of humanities and social sciences, be it archaeology, history, psychology and sociology to name but a few relevant fields. Researchers in these fields are not necessarily interested in language, *per se*, and certainly lack the required skills in language technology to develop themselves the infrastructure required for computational methods to be even considered. Every effort will be made not only to make resources available but to provide preferably off-the-shelf tools and solutions and the necessary training and advising to customize the resources in order to suit the particular needs of humanities researchers.